

An Algorithm for Tolerating Crash Failures in Distributed Systems

Vincenzo De Florio, Geert Deconinck, and Rudy Lauwereins

Katholieke Universiteit Leuven,
Electrical Engineering Department, ACCA Group,
Kard. Mercierlaan 94, B-3001 Heverlee, Belgium.

E-mail: {vincenzo.deflorio|lauwerin}@gmail.com

Abstract

In the framework of the ESPRIT project 28620 “TIRAN” (tailorable fault tolerance frameworks for embedded applications), a toolset of error detection, isolation, and recovery components is being designed to serve as a basic means for orchestrating application-level fault tolerance. These tools will be used either as stand-alone components or as the peripheral components of a distributed application, that we call “the backbone”. The backbone is to run in the background of the user application. Its objectives include (1) gathering and maintaining error detection information produced by TIRAN components like watchdog timers, trap handlers, or by external detection services working at kernel or driver level, and (2) using this information at error recovery time. In particular, those TIRAN tools related to error detection and fault masking will forward their deductions to the backbone that, in turn, will make use of this information to orchestrate error recovery, requesting recovery and reconfiguration actions to those tools related to error isolation and recovery. Clearly a key point in this approach is guaranteeing that the backbone itself tolerates internal and external faults. In this article we describe one of the means that are used within the TIRAN backbone to fulfill this goal: a distributed algorithm for tolerating crash failures triggered by faults affecting at most all but one of the components of the backbone or at most all but one of the nodes of the system. We call this the algorithm of mutual suspicion.

1. Introduction

In the framework of the ESPRIT project 28620 “TIRAN” [4], a toolset of error detection, isolation, and recovery components is being developed to serve as a basic means for orchestrating application-level software fault tolerance. The basic components of this toolset can be considered as ready-made software tools that the developer has

to embed into his/her application so to enhance its dependability. These tools include, e.g., watchdog timers, trap handlers, local and distributed voting tools.

The main difference between TIRAN and other libraries with similar purposes, e.g., ISIS [3] or HATS [12], is the adoption of a special component, located between the basic toolset and the user application. This entity is transparently replicated on each node of the system, to keep track of events originating in the basic layer (e.g., a missing heartbeat from a task guarded by a watchdog) or in the user application (e.g., the spawning of a new task), and to allow the orchestration of system-wide error recovery and reconfiguration. We call this component “the backbone”.

In order to perform these tasks, the backbone hooks to each active instance of the basic tools and is transparently informed of any error detection or fault masking event taking place in the system. Similarly, it also hooks to a library of basic services. This library includes, among others, functions for remote communication, for task creation and management, and to access the local hardware clock. These functions are instrumented so to transparently forward to the backbone notifications of events like the creation or the termination of a thread. Special low-level services at kernel or at driver-level are also hooked to the backbone—for example, on a custom board based on Analog Devices ADSP-21020 DSP and on IEEE 1355-compliant communication chips [1], communication faults are transparently notified to the backbone by driver-level tools. Doing this, an information stream flows on each node from different abstract layers to the local component of the backbone. This component maintains and updates this information in the form of a system database, also replicating it on different nodes.

Whenever an error is detected or a fault is masked, those TIRAN tools related to error detection and fault masking forward their deductions to the backbone that, in turn, makes use of this information to manage error recovery, requesting recovery and reconfiguration actions to those tools related to error isolation and recovery (see Fig. 1).

The specification of which actions to take is to be sup-

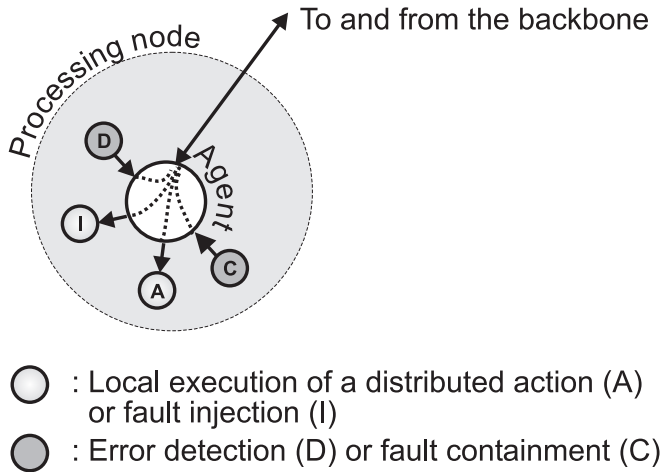


Figure 1. Each processing node of the system hosts one agent of the backbone. This component is the intermediary of the backbone on that node. In particular, it gathers information from the TIRAN tools for error detection and fault containment (grey circles) and forwards requests to those tools managing error containment and recovery (light grey circles). These latter execute recovery actions and possibly, at test time, fault injection requests.

plied by the user in the form of a “recovery script”, a sort of ancillary application context devoted to error recovery concerns, consisting of a number of guarded commands: the execution of blocks of basic recovery actions, e.g., restarting a group of tasks, rebooting a node and so forth, is then subject to the evaluation of boolean clauses based on the current contents of the system database—for further information on this subject see [9, 11].

Clearly a key point in this scheme is guaranteeing that the backbone itself tolerates internal as well external faults. In this article we describe one of the means that have been designed within the TIRAN backbone to fulfill this goal: a distributed algorithm for tolerating crash failures triggered by faults affecting at most all but one of the components of the backbone. We call this the algorithm of mutual suspicion. As in [6], we assume a timed asynchronous distributed system model [5]. Furthermore, we assume the availability of asynchronous communication means. No atomic broadcast primitive is required.

2. Basic assumptions

The target system is a distributed system consisting of n nodes ($n \geq 1$). Nodes are assumed to be labeled with a unique number in $\{0, \dots, n-1\}$. The backbone is a distributed application consisting of n triplets of tasks:

$$(\text{task } \mathcal{D}, \text{task } \mathcal{I}, \text{task } \mathcal{R}).$$

Basically task \mathcal{D} bears its name from the fact that it deals with the system database of the backbone, while task \mathcal{I} manages “I’m alive” signals, and task \mathcal{R} deals with error recovery (see further on). We call “agent” any such triplet. At initialisation time, on each node there is exactly one agent, identified through the label of the node where it runs on. For any $0 \leq k < n$, we call “ $t[k]$ ” task t of agent k . Agents play two roles: coordinator (also called manager) and assistant (also called backup agent or simply “backup”). In the initial, correct state there is just one coordinator. The choice of which node should host the coordinator, as well as system configuration and node labeling is done at compile time through a configuration script.

Figure 2 displays a backbone running on a four node system (a Parsytec Xplorer MIMD engine based on PowerPC microprocessors). In this case, node zero hosts the coordinator while nodes 1–3 execute as assistants. In the portrayed situation no errors have been detected, and this is rendered with green circles and a status label equal to “OK”.

Each agent, be it a coordinator or an assistant, executes a number of common tasks. Among these tasks we have:

1. Interfacing the instances of the basic tools of the framework—this is to be carried out by task \mathcal{D} .
2. Organizing/maintaining data gathered from the instances—also specific of task \mathcal{D} .
3. Error recovery and reconfiguration management (task \mathcal{R}).
4. Self-check. This takes place through a distributed algorithm that is executed by task \mathcal{D} and task \mathcal{I} in all agents in the case that $n > 1$.

This article does not cover points 1–3; in particular points 1 and 2 are dealt with as reported in [10], while the issue of error recovery and reconfiguration management is described in [9]. In the following, we describe the algorithm mentioned at point 4. As the key point of this algorithm is the fact that the coordinator and assistants mutually question their state, we call this the algorithm of mutual suspicion (AMS).



Figure 2. A Netscape browser offers a global view of the TIRAN backbone: number, role, and state of each component is displayed. “DIR net” is a nickname for the backbone. “OK” means that no faults have been detected. When the user selects a component, further information related to that component is displayed. The small icon on bottom links to a page with information related to error recovery.

3. The algorithm of mutual suspicion

This Section describes AMS. Simulation and testing show that AMS is capable of tolerating crash failures of up to $n - 1$ agents (some or all of which may be caused by a node crash). This implies fail-stop behaviour, that can be reached in hardware, e.g., by architectures based on duplication and comparison [14], and coupled with other techniques like, e.g., control flow monitoring [15] or diversification and comparison [13]. If a coordinator or its node crash, a non-crashed assistant becomes coordinator. Whenever a crashed agent is restarted, possibly after a node reboot, it is inserted again in the backbone as an assistant. Let us first sketch the structure of AMS. Let m be the node where the coordinator first runs on. In short:

- The coordinator periodically broadcasts a MIA (manager is alive) message as part of its task \mathcal{D} , the assistants periodically send the coordinator a TAIA (this as-

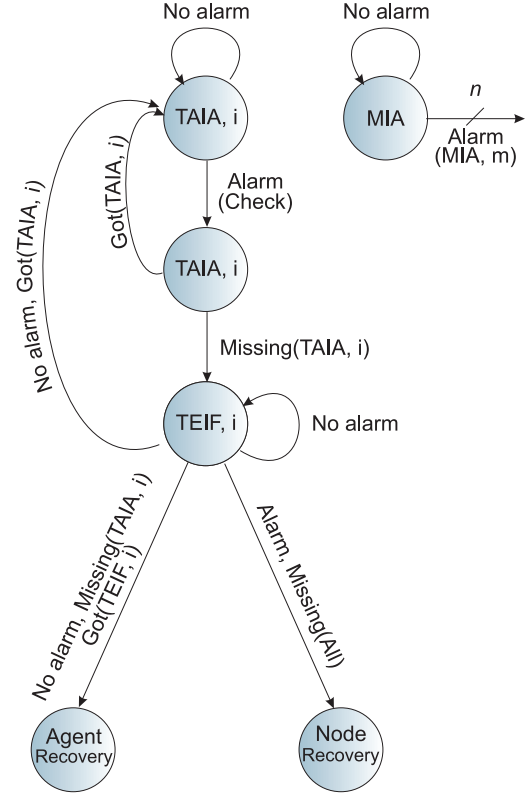


Figure 3. A representation of the algorithm of the manager.

sistant is alive) message as part of their task \mathcal{D} .

- For each $0 \leq k < n$, task $\mathcal{D}[k]$ periodically sets a flag. This flag is periodically cleared by task $\mathcal{I}[k]$.
- If, for any valid j , task $\mathcal{I}[j]$ finds that the flag has not been set during the period just elapsed, task $\mathcal{I}[j]$ broadcasts a TEIF (this entity is faulty) message.
- When any agent, say the agent on node g , does not receive any timely message from an other agent, say the agent on node f , be that a MIA or a TAIA message, then agent g enters a so-called suspicion-period by setting flag $\text{sus}[f]$. This state leads to three possible next states, corresponding to these events:
 1. Agent g receives a TEIF from task $\mathcal{I}[f]$ within a specific time period, say t clock ticks.
 2. Agent g receives a (late) TAIA from task $\mathcal{D}[f]$ within t clock ticks.
 3. No message is received by agent g from neither task $\mathcal{I}[f]$ nor task $\mathcal{D}[f]$ within t clock ticks.

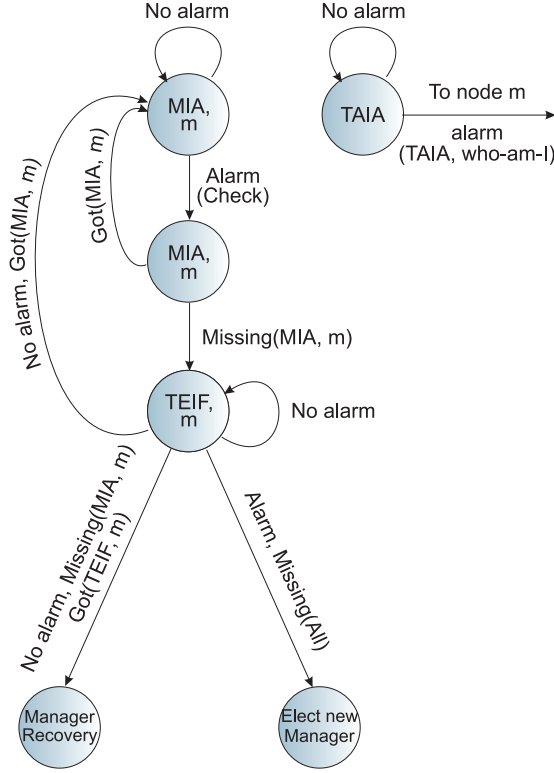


Figure 4. A representation of the algorithm of the assistant.

State 1 corresponds to deduction “agent on node f has crashed, though node f is still operational”. State 2 translates into deduction “both agent on node f and node f are operational, though for some reason agent f or its communication means have been slowed down”. State 3 is the detection of a crash failure for node f . These deductions lead to actions aiming at recovering agent f or (if possible) the whole node f , possibly electing a new coordinator. In the present version of AMS, the election algorithm is simply carried out assuming the next coordinator to be the assistant on node $m + 1 \bmod n$.

As compliant to the timed asynchronous distributed system model, we assume the presence of an alarm manager task (task \mathcal{A}) on each node of the system, spawned at initialization time by the agent. Task \mathcal{A} is used to translate time-related clauses, e.g., “ t clock ticks have elapsed”, into message arrivals. Let us call task $\mathcal{A}[j]$ the task \mathcal{A} running on node j , $0 \leq j < n$. Task \mathcal{A} may be represented as a function

$$a : C \rightarrow M$$

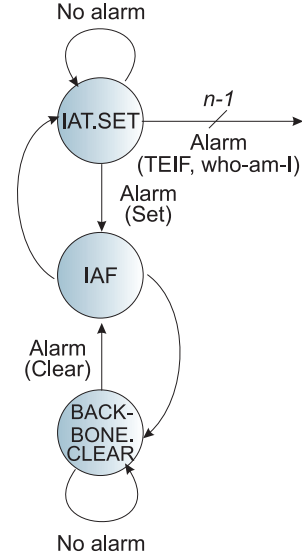


Figure 5. A representation of the conjoint action of task \mathcal{I} and task \mathcal{D} on the “I’m alive” flag.

such that, for any time-related clause $c \in C$,

$$a(c) = \text{message “clause } c \text{ has elapsed”} \in M.$$

Task $\mathcal{A}[j]$ monitors a set of time-related clauses and, each time one of them occurs, say clause c , it sends task $\mathcal{D}[j]$ message $a(c)$. Messages are sent via asynchronous primitives based on mailboxes.

In particular, the coordinator, which we assume to run initially on node m , instructs its task \mathcal{A} so that the following clauses be managed:

- (MIA_SEND, j , MIA_SEND.TIMEOUT), j different from m : every MIA_SEND.TIMEOUT clock ticks, a message of type (MIA_SEND, j) should be sent to task $\mathcal{D}[m]$, i.e., task \mathcal{D} on the current node. This latter will respond to such event by sending each task $\mathcal{D}[j]$ a MIA (manager is alive) message.
- (TAIA_RECV, j , TAIA_RECV.TIMEOUT) for each j different from m : every TAIA_RECV.TIMEOUT clock ticks at most, a message of type TAIA is to be received from task $\mathcal{D}[j]$. The arrival of such a message or of any other “sign of life” from task $\mathcal{D}[j]$ translates also in renewing the corresponding alarm. On the other hand, the arrival of a message of type (TAIA_RECV, k), k different from m , sent by task $\mathcal{A}[m]$, warns task $\mathcal{D}[m]$ that assistant on node k sent no sign of life throughout the TAIA_RECV.TIMEOUT-clock-tick period just elapsed. This makes task $\mathcal{D}[m]$ set its flag $\text{sus}[k]$.

```

coordinator()
{
    /* the alarm of these alarms simply sends a message of id
    <alarm-type>, subid = <subid> to the coordinator
    */
    insert alarms (IA-flag-alarm, MIA_SEND-alarms, TAIA_RECV-alarms)

    clear all entries of sus[] /* suspicion periods are all off */
    set IA-flag /* I'm alive flag */
    activate IAT /* I'm Alive Task */

    loop (wait for incoming messages)
    {
        switch (message.type)
        {
            /* time to set the IA-flag! */
            case IA-flag-alarm:
                set IA-flag,
                break;

            /* time to send a MIA to an assistant */
            case MIA_SEND-alarm:
                send MIA to assistant <subid>
                break;

            /* a message which modifies the db */
            case DB:
                if (local)
                {
                    renew MIA_SEND alarm on all <subid>s
                    broadcast modifications to db
                    break;
                }
                else /* if it's a remote message, it's also
                a piggybacked TAIA */
                {
                    update your copy of the db
                    /* don't break */
                }

            /* if a TAIA comes in or a remote DB message comes in... */
            case TAIA:
                if ( ! ispresent(TAIA_RECV-alarm, <subid> ) )
                {
                    insert TAIA_RECV-alarm, <subid>
                    broadcast NIUA /* node is up again! */
                }
                renew TAIA_RECV-alarm, <subid>
                /* if you get a TAIA while expecting a TEIF, then
                no problem, simply get out of the suspicion period
                */
                if (sus[ <subid> ] == TRUE)
                {
                    sus[ <subid> ] = FALSE;
                    break;
                }

            /* no heartbeat from a remote component... enter
            a suspicion period then
            */
            case TAIA_RECV-alarm:
                sus[ <subid> ] = TRUE
                insert alarm (TEIF_RECV-alarm, NON_CYCLIC,
                IMALIVE_RECV_TIMEOUT, <subid>)
                delete alarm (TAIA_RECV-alarm, <subid>);
                break;

            /* a TEIF message has been sent from a IAT:
            as its last action, the IAT went to sleep
            */
            case TEIF:
                if (sus[ <subid> ] == TRUE)
                {
                    delete alarm (TEIF_RECV-alarm, <subid>)
                    sus[ <subid> ] = FALSE
                    /* agent recovery will spawn a clone of the
                    assistant <subid>. If no assistant clones
                    are available, the entire node will
                    be rebooted.
                    */
                    Agent-Recovery( <subid> )
                }
                else
                {
                    if (local)
                    {
                        set IA-flag
                        enable IAT
                    }
                    else
                    {
                        send ENIA (i.e., "ENable IAT") to <subid>
                    }
                }
                break;

            case TEIF_RECV-alarm:
                if (sus[ <subid> ] == TRUE)
                {
                    delete alarm (TAIA_RECV-alarm, <subid>)
                    sus[ <subid> ] = FALSE
                    /* the entire node will be rebooted.
                    */
                    Node-Recovery( <subid> )
                }
                break;

            case ENIA:
                set IA-flag
                activate IAT
                /* if an IAT gets an "activate" message while it's
                active, the message is ignored */
                break;

            default:
                deal with these other messages
        } /* end switch (message.type) */

        set IA-flag
        renew IA-flag-alarm
    } /* end loop */
} /* end coordinator */

```

Figure 6. Pseudo-code of the coordinator.

- (I'M_ALIVE_SET, m , I'M_ALIVE_SET.TIMEOUT): every I'M_ALIVE_SET.TIMEOUT clock ticks, task $\mathcal{A}[m]$ sends task $\mathcal{D}[m]$ an I'M_ALIVE_SET message. As a response to this, task $\mathcal{D}[m]$ sets the “I’m alive” flag, a memory variable that is shared between tasks of type \mathcal{D} and \mathcal{I} .

Furthermore, whenever flag $\text{sus}[k]$ is set, for any k different from m , the following clause is sent to task \mathcal{A} for being managed:

- (TEIF_RECV, k , TEIF_RECV.TIMEOUT): this clause simply asks task $\mathcal{A}[m]$ to schedule the sending of message (TEIF_RECV, k) to task $\mathcal{D}[m]$ after TEIF_RECV.TIMEOUT clock ticks. This action is canceled should a late TAIA message arrive to task $\mathcal{D}[m]$ from task k , or should a TEIF message from task $\mathcal{I}[k]$ arrive instead. In the first case, $\text{sus}[k]$ is cleared and (possibly empty) actions corresponding to a slowed down task $\mathcal{D}[k]$ are taken. In the latter case, task $\mathcal{D}[k]$ is assumed to have crashed, its clauses are removed from the list of those managed by task $\mathcal{A}[m]$, and flag $\text{sus}[k]$ is cleared. It is assumed that task $\mathcal{I}[k]$ will take care in this case of reviving task $\mathcal{D}[k]$. Any future sign of life from task $\mathcal{D}[k]$ is assumed to mean that task $\mathcal{D}[k]$ is back in operation. In such a case task $\mathcal{D}[k]$ would then be re-entered in the list of operational assistants, and task $\mathcal{D}[m]$ would then request task $\mathcal{A}[m]$ to include again an alarm of type (MIA_SEND, k , MIA_SEND.TIMEOUT) and an alarm of type (TAIA_RECV, k , TAIA_RECV.TIMEOUT) in its list. If a (TEIF_RECV, k) message reaches task $\mathcal{D}[m]$, the entire node k is assumed to have crashed. Node recovery may start at this point, if available, or a warning message should be sent to an external operator so that, e.g., node k be rebooted.

Similarly any assistant, say the one on node k , instructs its task \mathcal{A} so that the following clauses be managed:

- (TAIA_SEND, m , TAIA_SEND.TIMEOUT): every TAIA_SEND.TIMEOUT clock ticks, a message of type (TAIA_SEND, m) is to be sent to task $\mathcal{D}[k]$, i.e., task \mathcal{D} on the current node. This latter will respond to such event by sending task $\mathcal{D}[m]$ (i.e., the manager) a TAIA (this agent is alive) message. Should a data message be sent to the manager in the middle of TAIA_SEND.TIMEOUT-clock-tick period, alarm (TAIA_SEND, m , TAIA_SEND.TIMEOUT) is renewed. This may happen for instance because one of the basic TIRAN tools for error detection reports an event to task $\mathcal{D}[k]$. Such event must be sent to the manager for it to update its database. In this case we say that the TAIA message is sent in piggybacking with the event notification message.

```

assistant()
{
    /* the alarm of these alarms simply sends a message of id
    <alarm-type>, subid = <subid> to the assistant agent
    */
    insert alarms (IA-flag-alarm, MIA_RECV-alarm, TAIA_SEND-alarm)
    /* only one MIA and one TAIA alarm */

    clear sus          /* suspicion period is set to off */
    clear IA-flag      /* 'I'm alive' flag */
    activate IAT       /* I'm Alive Task */

    loop (wait for incoming messages)
    {
        switch (message.type)
        {
            /* time to set the IA-flag */
            case IA-flag-alarm:
                set IA-flag,
                break;

            /* time to send a TAIA to the coordinator */
            case TAIA-alarm:
                send (TAIA, <subid=MyprocID>) to <coordinator>
                break;

            /* if a MIA comes in... */
            case MIA:
                if ( ! ispresent (MIA_RECV-alarm, <subid> ) )
                {
                    insert MIA_RECV-alarm, <subid>
                    /* a new coordinator has been chosen */
                    <coordinator> = <subid>
                }

                renew MIA_RECV-alarm, <coordinator>
                /* if you get a MIA while expecting a TEIF, then
                no problem, simply get out of the suspicion period
                */
                if (sus == TRUE)
                    sus = FALSE;
                break;

            /* no heartbeat from the coordinator... enter
            a suspicion period then
            */
            case MIA_RECV-alarm:
                sus = TRUE
                insert alarm (TEIF_RECV-alarm, NON_CYCLIC,
                             IMALIVE_RECV_TIMEOUT, <coordinator>)
                delete alarm (MIA_RECV-alarm, <subid>);
                break;

            /* a TEIF message has been sent from a IAT:
            as its last action, the IAT went to sleep
            */
            case TEIF:
                if (sus == TRUE)
                {
                    delete alarm (TEIF_RECV-alarm, <subid>)
                    sus = FALSE

                    /* Manager recovery will spawn a clone of the
                    assistant <subid>. If no more coordinator clones
                    are available, the entire node will
                    be rebooted.
                    */
                    Manager-Recovery( <coordinator> )
                }
                else
                {
                    send ENIA (i.e., 'ENable IAT') to <coordinator>
                    renew MIA_RECV-alarm
                }
                break;

            case TEIF_RECV-alarm:
                if (sus == TRUE)
                {
                    delete alarm (MIA_RECV-alarm, <subid>)
                    sus = FALSE
                    /* the entire node will be rebooted.
                    */
                    Node-Recovery( <subid> )
                    send ANID to all except <coordinator>
                    choose next coordinator
                    if this assistant is to be the new coordinator
                    {
                        mystate->configuration = MANAGER;
                        SetState(&mystate);
                        RestartTask();
                    }
                }
                else
                {
                    send ENIA (i.e., 'ENable IAT') to <coordinator>
                    renew MIA_RECV-alarm
                }
                break;

            case ENIA:
                set IA-flag
                activate IAT
                /* if an IAT gets an 'activate' message while it's
                active, the message is ignored */
                break;

            default:
                deal with these other messages
        } /* end switch (message.type) */

        set IA-flag
        renew IA-flag-alarm
    } /* end loop */
} /* end assistant */

```

Figure 7. Pseudo-code of the assistant.

- (MIA_RECV, m , MIA_RECV_TIMEOUT): every MIA_RECV_TIMEOUT clock ticks at most, a message of type MIA is to be received from task $\mathcal{D}[m]$, i.e., the manager. The arrival of such a message or of any other “sign of life” from the manager translates also in renewing the corresponding alarm. If a message of type (MIA_RECV, m), is received from task $\mathcal{D}[k]$ and sent by task $\mathcal{A}[k]$, this means that no sign of life has been received from the manager throughout the MIA_RECV_TIMEOUT-clock-tick period just elapsed. This makes task $\mathcal{D}[k]$ set flag $\text{sus}[m]$.
- (I'M_ALIVE_SET, k , I'M_ALIVE_SET_TIMEOUT): every I'M_ALIVE_SET_TIMEOUT clock ticks, task $\mathcal{A}[k]$ sends task $\mathcal{D}[k]$ an I'M_ALIVE_SET message. As a response to this, task $\mathcal{D}[k]$ sets the “I’m alive” flag.

Furthermore, whenever flag $\text{sus}[m]$ is set, the following clause is sent to task $\mathcal{A}[k]$ for being managed:

- (TEIF_RECV, m , TEIF_RECV_TIMEOUT): this clause simply asks task $\mathcal{A}[k]$ to postpone sending message (TEIF_RECV, k) to task $\mathcal{D}[k]$ of TEIF_RECV_TIMEOUT clock ticks. This action is canceled should a late MIA message arrive to task $\mathcal{D}[k]$ from the manager, or should a TEIF message from task $\mathcal{I}[m]$ arrive instead. In the first case, $\text{sus}[m]$ is cleared and possibly empty actions corresponding to a slowed down manager are taken. In the latter case, task $\mathcal{D}[m]$ is assumed to have crashed, its clause is removed from the list of those managed by task $\mathcal{A}[k]$, and flag $\text{sus}[m]$ is cleared. It is assumed that task $\mathcal{I}[m]$ will take care in this case of reviving task $\mathcal{D}[m]$. Any future sign of life from task $\mathcal{D}[m]$ is assumed to mean that task $\mathcal{D}[m]$ is back in operation. In such a case task $\mathcal{D}[m]$ would be demoted to the role of assistant and entered in the list of operational assistants. The role of coordinator would then have been assigned, via an election, to an agent formerly running as assistant.

If a (TEIF_RECV, m) message reaches task $\mathcal{D}[k]$, the entire node of the manager is assumed to have crashed. Node recovery may start at this point. An election takes place—the next assistant (modulo n) is elected as new manager.

Also task \mathcal{I} on each node, say node k , instructs its task \mathcal{A} so that the following clause be managed:

- (I'M_ALIVE_CLEAR, k , I'M_ALIVE_CLEAR_TIMEOUT): every I'M_ALIVE_CLEAR_TIMEOUT clock ticks task $\mathcal{A}[k]$ sends task $\mathcal{I}[k]$ an I'M_ALIVE_CLEAR message. As a response to this, task $\mathcal{I}[k]$ clears the “I’m alive” flag.

Figures 3, 4, and 5 supply a pictorial representation of this algorithm. Figure 6 and Fig. 7 respectively show a pseudo-code of the coordinator and of the assistant.

3.1. The alarm manager class

This section briefly describes task \mathcal{A} . This task makes use of a special class to manage lists of alarms [8]. The class allows the client to “register” alarms, specifying alarm-ids and deadlines.

Once the first alarm is entered, the task managing alarms creates a linked-list of alarms and polls the top of the list. For each new alarm to be inserted, an entry in the list is found and the list is modified accordingly. If the top entry expires, a user-defined alarm function is invoked. This is a general mechanism that allows to associate any event with the expiring of an alarm. In the case of the backbone, task \mathcal{A} on node k sends a message to task $\mathcal{D}[k]$ —the same result may also be achieved by sending an UNIX signal to task $\mathcal{D}[k]$. Special alarms are defined as “cyclic”, i.e., they are automatically renewed at each new expiration, after invoking the alarm function. A special function restarts an alarm, i.e., it deletes and re-enters an entry. It is also possible to temporarily suspend an alarm and re-enable it afterwards.

4. Current status and future directions

A prototypal implementation of the TIRAN backbone is running on a Parsytec Xplorer, a MIMD engine, using 4 PowerPC nodes. The system has been tested and proved to be able to tolerate a number of software-injected faults, e.g., component and node crashes (see Fig. 8 and Fig. 9). Faults are scheduled as another class of alarms that, when triggered, send a fault injection message to the local task \mathcal{D} or task \mathcal{I} . The specification of which fault to inject is read by the backbone at initialisation time from a file called “faultrc”. The user can specify fault injections by editing this file, e.g., as follows:

```
INJECT CRASH ON COMPONENT 1
    AFTER 5000000 TICKS
INJECT CRASH ON NODE 0
    AFTER 10000000 TICKS.
```

The first two lines inject a crash on task $\mathcal{D}[1]$ after 5 seconds from the initialisation of the backbone, the second ones inject a system reboot of node 0 after 10 seconds. Via fault injection it is also possible to slow down artificially a component for a given period. Slowed down components are temporarily and automatically disconnected and then accepted again in the application when their performance goes back to normal values. Scenarios are represented into a Netscape window where a monitoring application displays the structure of the user application, maps the backbone roles onto the processing nodes of the system, and constantly reports about the events taking place in the system [7].

This system has been recently redeveloped so to enhance its portability and performance and to improve its

resilience. The backbone is currently being implemented for target platforms based on Windows CE, VxWorks, and TEX [2]. A GSPN model of the algorithm of mutual suspicion has been developed by the University of Turin, Italy, and has been used to validate and evaluate the system. Simulations of this model proved the absence of deadlocks and livelocks. Measurements of the overheads in fault free scenarios and when faults occur will also be collected and analysed.

Acknowledgments. This project is partly supported by an FWO Krediet aan Navorsers and by the ESPRIT-IV project 28620 “TIRAN”.

References

- [1] Anonymous. IEEE standard for Heterogeneous Interconnect (HIC) (Low-cost, low-latency scalable serial interconnect for parallel system construction). Technical Report 1355-1995 (ISO/IEC 14575), IEEE, 1995.
- [2] Anonymous. *TEX User Manual*. TXT Ingegneria Informatica, Milano, Italy, 1997.
- [3] K. P. Birman. Replication and fault tolerance in the Isis system. *ACM Operating Systems Review*, 19(5):79–86, 1985.
- [4] Oliver Botti, Vincenzo De Florio, Geert Deconinck, Susanna Donatelli, Andrea Bobbio, Axel Klein, H. Kufner, Rudy Lauwereins, E. Thurner, and E. Verhulst. TIRAN: Flexible and portable fault tolerance solutions for cost effective dependable applications. In P. Amestoy et al., editors, *Proc. of the 5th Int. Euro-Par Conference (EuroPar’99), Lecture Notes in Computer Science*, volume 1685, pages 1166–1170, Toulouse, France, August/September 1999. Springer-Verlag, Berlin.
- [5] Flaviu Cristian and Christof Fetzer. The timed asynchronous distributed system model. *IEEE Trans. on Parallel and Distributed Systems*, 10(6):642–657, June 1999.
- [6] Flaviu Cristian and Frank Schmuck. Agreeing on processor group membership in asynchronous distributed systems. Technical Report CSE95-428, UCSD, 1995.
- [7] Vincenzo De Florio, Geert Deconinck, Mario Truyens, Wim Rosseel and Rudy Lauwereins. A hypermedia distributed application for monitoring and fault-injection in embedded fault-tolerant parallel programs. In *Proc. of the 6th Euromicro Workshop on Parallel and Distributed Processing (PDP’98)*, pages 349–355, Madrid, Spain, January 1998. IEEE Comp. Soc. Press.

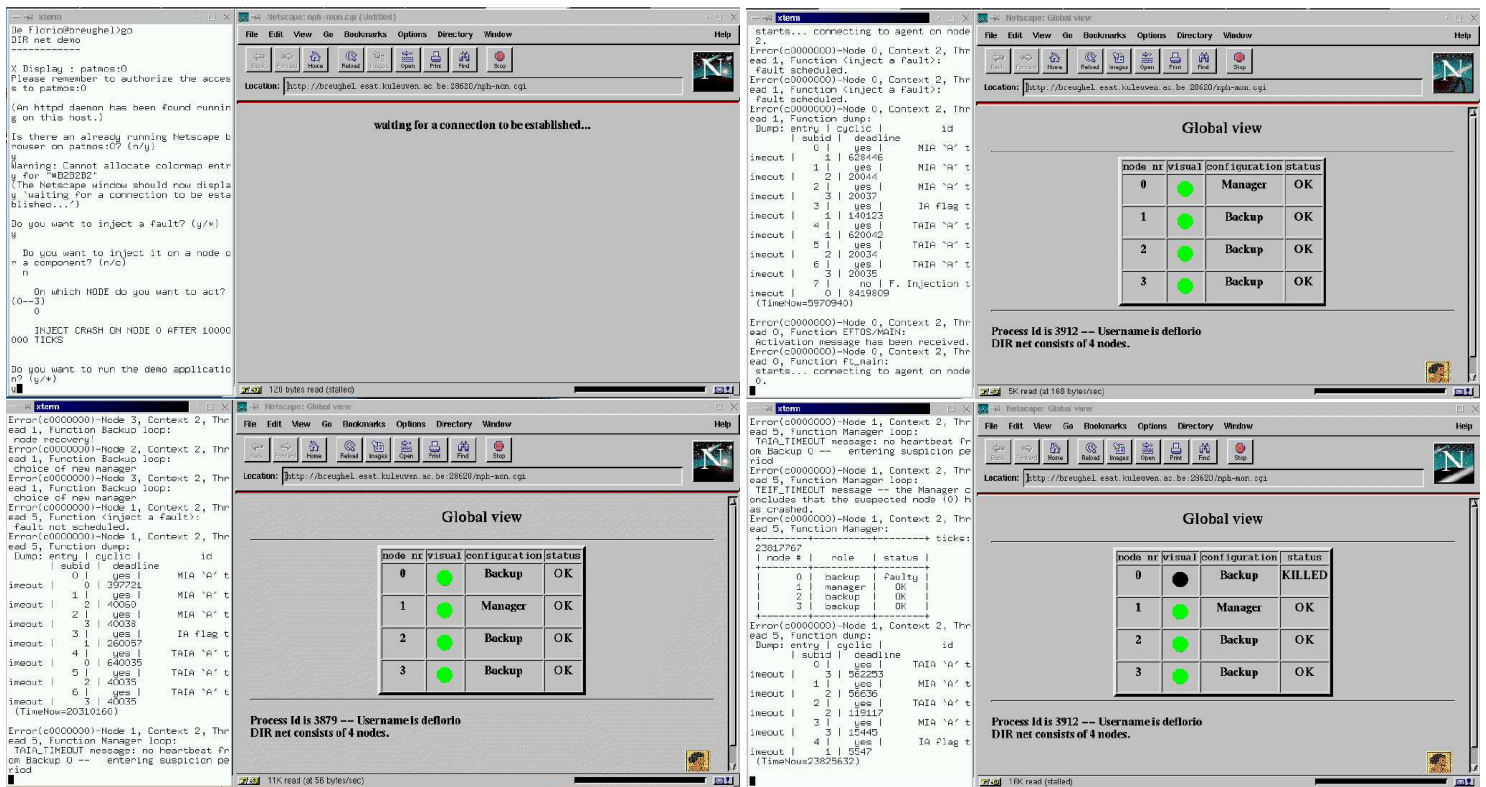


Figure 8. A fault is injected on node 0 of a system of four nodes. Node 0 hosts the manager of the backbone. In the top left picture the user selects the fault to be injected and connects to a remotely controllable Netscape browser. The top right picture shows this latter as it renders the shape and state of the system. The textual window reports the current contents of the list of alarms used by task $A[0]$. In the bottom left picture the crash of node 0 has been detected and a new manager has been elected. On election, the manager expects node 0 to be back in operation after a recovery step. This recovery step is not performed in this case. As a consequence, node 0 is detected as inactive and labeled as “KILLED”.

- [8] Vincenzo De Florio, Geert Deconinck, and Rudy Lauwereins. A time-out management system for real-time distributed applications. Submitted for publication in *IEEE Trans. on Computers*.
- [9] Vincenzo De Florio, Geert Deconinck, and Rudy Lauwereins. The recovery language approach for software-implemented fault tolerance. Submitted for publication in *ACM Transactions on Computer Systems*.
- [10] Geert Deconinck, Vincenzo De Florio, Rudy Lauwereins, and Ronnie Belmans. A software library, a control backbone and user-specified recovery strategies to enhance the dependability of embedded systems. In *Proc. of the 25th Euromicro Conference (Euromicro '99), Workshop on Dependable Computing Sys-*

tems, volume 2, pages 98–104, Milan, Italy, September 1999. IEEE Comp. Soc. Press.

- [11] Geert Deconinck, Mario Truyens, Vincenzo De Florio, Wim Rosseel, Rudy Lauwereins, and Ronnie Belmans. A framework backbone for software fault tolerance in embedded parallel applications. In *Proc. of the 7th Euromicro Workshop on Parallel and Distributed Processing (PDP'99)*, pages 189–195, Funchal, Portugal, February 1999. IEEE Comp. Soc. Press.
- [12] Yennun Huang and Chandra M.R. Kintala. Software fault tolerance in the application layer. In Michael Lyu, editor, *Software Fault Tolerance*, chapter 10, pages 231–248. John Wiley & Sons, New York, 1995.
- [13] D. Powell. Preliminary definition of the GUARDS architecture. Technical Report 96277, LAAS-CNRS, January 1997.

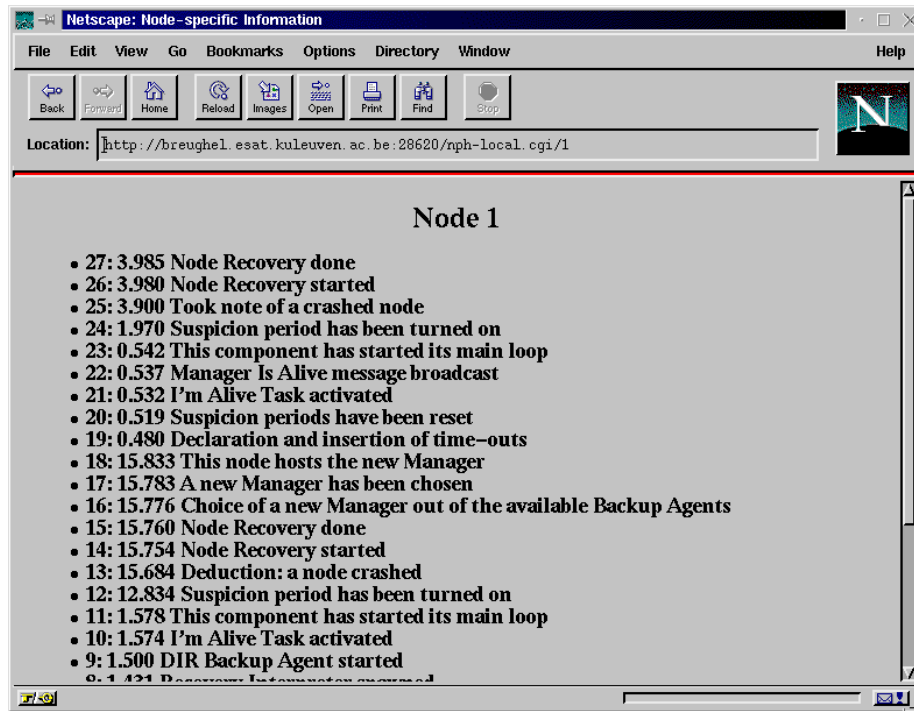


Figure 9. When the user selects a circular icon in the Web page of Fig. 2, the browser replies with a listing of all the events that took place on the corresponding node. Here, a list of events occurred on task $\mathcal{D}[1]$ during the experiment shown in Fig. 8 is displayed. Events are labeled with an event-id and with the time of occurrence (in seconds). Note in particular event 15, corresponding to deduction “a node has crashed”, and events 16–18, in which the election of the manager takes place and task $\mathcal{D}[1]$ takes over the role of the former manager, task $\mathcal{D}[0]$. Restarting as manager, task $\mathcal{D}[1]$ resets the local clock.

- [14] D. K. Pradhan. *Fault-Tolerant Computer Systems Design*. Prentice-Hall, Upper Saddle River, NJ, 1996.
- [15] M. Schuette and J. P. Shen. Processor control flow monitoring using signed instruction streams. *IEEE Trans. on Computers*, 36(3):264–276, March 1987.